Valuing Enterprise Data: CLDS Research Project Brief

James Short, Ph.D.
Lead Scientist, Center for Large Scale Data Systems (CLDS)
San Diego Supercomputer Center
University of California, San Diego
What Happens in an Internet Minute?

- 639,800 GB of global IP data transferred
- 20 New victims of identity theft
- 204 million Emails sent
- 47,000 App downloads
- 583,000 New LinkedIn accounts
- 61,141 Hours of music
- 20 million Photo views
- 320+ New Twitter accounts
- 3,000 Photo uploads
- 100,000 New tweets
- 135 Botnet infections
- 1,300 New mobile users
- 277,000 Logins
- 2.5 million Facebook views
- 2+ million Search queries
- 30 Hours of video uploaded
- 1.3 million Video views

And Future Growth is Staggering

Today, the number of networked devices
= the global population

By 2015, the number of networked devices
= 2x the global population

In 2015, it would take you 5 years
to view all video crossing IP networks each second
How Much Media?

UCSD–SDSC News, Nov 6, 2013

U.S. Media Consumption to Rise to 15.5 Hours a Day – Per Person – by 2015

- New Study Issued by SDSC Researcher with the USC Marshall School of Business
What is the VALUE in Big Data?

• **H$_1$**: It’s obvious (correlated with growth)
  • More Data = More Can Be Done = More Value

• **H$_2$**: It’s contingent on cost-effective use
  • It’s what you do with it and how effectively you do it
  • Value = ABVR + EFBV - Cost - Risk

• **H$_3$**: It’s a lot more than just more data and data activity
  • Operational performance, market expansion, customer intimacy, business scope
  • Valuation may finally hit the balance sheet
CLDS Project on Data Value

• Define a framework and processes for data valuation to improve business performance

• Project segments:
  • Build and validate model of data value
  • Identify key use cases and conduct field studies
  • Develop industry survey of valuation practices
  • Analyze field data against valuation model
  • Iterate with project sponsors and case study participants
  • Come together at an SDSC Symposium in the 3rd or 4th quarter of this year
CLDS Project on Data Value

- Valuation model
  - Components: Asset – Activity – Expected Value
  - Model refined and validated in use cases and surveys of valuation practices

- Valuation Practices
  - Management practices (data and information governance)
  - IT systems (rules embedded in software, automated)
  - Regulatory activities (rules enacted by government, etc.)
  - Merger and acquisition valuation
  - Classification of practices:
    - Develop a taxonomy of practices and analyze practices and their relationships
Presentation Summary

- Provide a brief background of SDSC, CLDS, and the data value project
- Outline the key questions and what we hope to accomplish on the project
- Present overview of the valuation model and discuss examples
- Present and discuss Facebook data
- Outline next steps and how you can get involved
- Q&A
**Brief History of SDSC**

- **1985-1997**: NSF national supercomputer center; managed by General Atomics
  - NSF PACI program leadership center; managed by UCSD
    - PACI: Partnerships for Advanced Computational Infrastructure
- **1997-2007**: NSF PACI program leadership center; managed by UCSD
- **2007-2009**: Internal transition to support more diversified research computing
  - NSF national “resource provider”
- **2009-future**: Multi-constituency cyberinfrastructure (CI) center
  - provide data-intensive CI resources, services, and expertise for university and industry research, CA and US projects and programs (ex: NIH)
- **Approaching $1B in lifetime contract and grant activity**
• A “data-intensive” supercomputer based on SSD flash memory and virtual shared memory
  • Emphasizes MEM and IO over FLOPS
• Scalable Data Mining
• Visualizations
CLDS: Center for Large Scale Data Systems Research

• Research center focused on the technical, business and economic challenges of next era compute, storage and network architectures

• Technical systems research
  • Workshop on Big Data Benchmarking

• Technology management research
  • Data value research program
    • Focused on the business value of data and information
  • How Much Information?
    • Focused on the rate of data and information growth
CLDS: Center for Large Scale Data Systems Research

• Support from government and industry
  • National Science Foundation (NSF)
  • National Institute of Standards and Technology (NIST)

• Industry Forum
  • Lead Center Sponsors
    • Seagate, Pivotal, NetApp, Intel, Brocade
  • Participating Members
    • Infosys, Mellanox, Persistent Systems, SAP
What is so hard about defining value?

• What is value?
  • Tangible, Intangible
  • Utility (rational use), Engagement (activity), Subjective

• Context
  • Value to whom?
  • Intrinsic (art), Extrinsic (defined in use, in exchange)
  • Parties to exchange: No assumption of = value

• Time
  • Value decays over time; value increases over time
  • Activity value erodes over time. What about stored data?
  • Econ 100: The value of any asset erodes without maintenance expenditures
What is so hard about defining value?

- Old MIT Media Lab example (Nick Negroponte)
- Doesn’t appear on corporate balance sheets or 10K disclosures
- If any business can sort it out, shouldn’t that be “information services firms”?
  - **Netflix**: Makes its money on subscription fees and advertising. The value of its information assets (content library) is expressed as the cost of licensing or creating that content (following traditional accounting rules for intangibles)
  - **Bloomberg**: Its core business is its terminal operations. It charges subscribers user fees, access fees (to information databases) and data services fees (analytics)
- Is there a common pattern here?
**Bottom Line**

- “Value” only has practical meaning when the parties concerned have been identified and the need for valuation spelled out
- Tangible assets are accounted for
- “Intangible” assets are a mix of formal rules and judgments
  - Cost of creation
    - *How much for someone else to duplicate it today?*
  - Capitalization of income
    - *Estimated future benefit accruing from ownership of the asset*
  - Discounted cash flow
    - *Method for gauging the attractiveness of an investment opportunity*
Starting Point

- **Value** \((V^{Act}) = Data\ Activity\)
  - H1: the greater the activity on data, the greater the value
    - Where value is defined by the number of users requesting the data, and the resources required by applications using it
    - Benefits: Intuitive, compartmentalized, measurable, uniform, engineering driven
    - Drawbacks: Intuitive, engineering driven (business value may be far lower or higher), easy to think of use cases which do not follow the assumptions (and in these cases will mis-specify value)
Value = EV^{store} + V^{act} \text{ where activity = compute + network}

H1: V^{t} = Decreasing Returns to Scale
Total Value Output increases by less than the proportional change in inputs

H2: V^{t} = Increasing Returns to Scale
Total Value Output increases by more than the proportional change in inputs

- Base Transaction Activity Increasing by 50% Each Period
- Information Generated Increasing by 100% Each Period
- Retained (Stored) Knowledge Increasing by 25% Each Period
FACEBOOK EXAMPLE
Facebook Storage Systems

- **Online Transaction Processing Databases (OLTP)**
  - The Facebook Social Graph
- **Semi-online Light Transaction Processing Databases (SLTP)**
  - Facebook Messages and Facebook Time Series
- **Immutable DataStore**
  - Photos, videos, etc
- **Analytics DataStore**
  - Data Warehouse, Logs storage

## Size and Scale of Databases

<table>
<thead>
<tr>
<th>Database Type</th>
<th>Total Size</th>
<th>Technology</th>
<th>Bottlenecks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook Graph</td>
<td>Single digit petabytes</td>
<td>MySQL and TAO</td>
<td>Random read IOPS</td>
</tr>
<tr>
<td>Facebook Messages and Time Series Data</td>
<td>Tens of petabytes</td>
<td>HBase and HDFS</td>
<td>Write IOPS and storage capacity</td>
</tr>
<tr>
<td>Facebook Photos</td>
<td>High tens of petabytes</td>
<td>Haystack</td>
<td>storage capacity</td>
</tr>
<tr>
<td>Data Warehouse</td>
<td>Hundreds of petabytes</td>
<td>Hive, HDFS and Hadoop</td>
<td>storage capacity</td>
</tr>
</tbody>
</table>

# Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Query Latency</th>
<th>Consistency</th>
<th>Durability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook Graph</td>
<td>&lt; few millisec</td>
<td>quickly consistent across data centers</td>
<td>No data loss</td>
</tr>
<tr>
<td>Facebook Messages and Time Series Data</td>
<td>&lt; 200 millisec</td>
<td>consistent within a data center</td>
<td>No data loss</td>
</tr>
<tr>
<td>Facebook Photos</td>
<td>&lt; 250 millisec</td>
<td>immutable</td>
<td>No data loss</td>
</tr>
<tr>
<td>Data Warehouse</td>
<td>&lt; 1 min</td>
<td>not consistent across data centers</td>
<td>No silent data loss</td>
</tr>
</tbody>
</table>

Size and Scale of Messages Database

- 6 Billion messages/day
- 74 Billion operations/day
- At peak: 1.5 million operations/sec
- 55% read, 45% write operations
- Average write operation inserts 16 records
- All data is lzo compressed
- Growing at 8 TB/day

# Facebook Photo DataStore

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2012</th>
<th>Est Rate of Change (CAGR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Size</td>
<td>15 billion photos 1.5 Petabyte</td>
<td>High tens of petabytes</td>
<td>150%</td>
</tr>
<tr>
<td>Upload Rate</td>
<td>30 million photos/day 3 TB/day</td>
<td>300 million photos/day 30 TB/day</td>
<td>78%</td>
</tr>
<tr>
<td>Serving Rate</td>
<td>555K images/sec</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

# Analytics Data Growth (Last Four Years)

<table>
<thead>
<tr>
<th></th>
<th>Facebook Users</th>
<th>Queries/Day</th>
<th>Scribe Data/Day</th>
<th>Nodes in warehouse</th>
<th>Size (Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth</td>
<td>14X</td>
<td>60X</td>
<td>250X</td>
<td>260X</td>
<td>2500X</td>
</tr>
</tbody>
</table>

*If these growth factors are applied to estimated data in each category, what would the totals look like...

<table>
<thead>
<tr>
<th>Year</th>
<th>Facebook Users</th>
<th>Queries/Day</th>
<th>Scribe Data/Day</th>
<th>Nodes in warehouse</th>
<th>Size (Total)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>100 million*</td>
<td>1,000,000*</td>
<td>5,000,000*</td>
<td>350*</td>
<td>100 TB*</td>
</tr>
<tr>
<td>2012</td>
<td>1.4 billion*</td>
<td>60,000,000*</td>
<td>1,250,000,000*</td>
<td>91,000*</td>
<td>250 PB*</td>
</tr>
</tbody>
</table>

*Stylized values (not FB data). For illustrative purposes only.
Summary Points

• Finding value in FB rate of growth is both a huge opportunity and challenge
• FB reports storage capacity bottlenecks in data warehouse and photo/video DataStore (again a function of growth rates)
• Value of DataStore not on FB balance sheet (nor are any of FB’s data assets for that matter)
Invitation to Join the Project

• The DV project is open to interested individuals and companies who would like to get involved

• Engagement can mean one or more of several things:
  • As a project sponsor and member of the CLDS center
  • As a project participant and Symposium sponsor
  • As a use case / field study participant
  • Later in the year, as a member of the project network
Project Goals

• To identify, contrast and compare the state of art practices in data and information valuation
• To develop a valuation taxonomy, conduct field studies and classify use cases according to the taxonomy
• To contribute to company practices in valuation, including descriptions and analyses, and developing experimental software tools
• In the second phase of the project, to validate the model and key conclusions by collecting and analyzing valuation data from an SDSC “big data” operational system
**Project Deliverables**

- A field-based model of data and information valuation
- A survey and analysis of corporate practices in valuation
- Use cases and field studies (“lessons learned”)
- Experimental software tools
- Milestone report backs and a Symposium on Data Value, to be held at SDSC late this year

- DV is designed as a 20 month project
  - First phase model definition and development concluded in December 2013
  - First project phase started Jan 1, 2014
Next Steps

- Interested parties should get in contact with PI Jim Short (jshort@ucsd.edu)
- Next update briefing on the project will be in June
- The CLDS website is currently being redesigned. Relaunch in March.
Questions?

- [www.sdsc.edu](http://www.sdsc.edu)
- For further information, contact Jim Short (jshort@ucsd.edu)

Thank you!